Data Intensive Scientific Computing on Petabyte Scalable Infrastructure, Phase I

Completed Technology Project (2009 - 2009)

Project Introduction

The infrastructure and programming paradigm for petabyte-level data processing performed at companies like Google and Yahoo shed some promising lights on the data-intensive scientific computing. Open source software and inexpensive commodity hardware make proprietary technologies within the grasp of academic communities. By leveraging these commercially proven and publicly available technologies, we are going to develop a suite of novel data management and analysis libraries, as an extension to existing primitive algorithms originally designed for web search. These libraries take advantage of the underlying petabyte-scalable data infrastructure, parallelize computation transparently and allow scientists and future commercial users to perform rather complex tasks (data mining, data visualization and machine learning) in a data intensive environment.

Anticipated Benefits

Data-intensive computing is not a problem unique to IT companies like Google. Nowadays, infrastructure and data analysis tools to support Data-Intensive-Scalable-Computing (DISC) are becoming competitive advantage even for non-IT companies, so that they can roll out new products and services faster and cheaper. For example, Wal-Mart sells ~300 million items everyday at 6000 stores worldwide. The entire data warehouse to support its business is as large as 4 PB. Scalable and efficient data analysis tool is vital to manage its supply chain, conduct market trend analysis and devise pricing strategy. A simple data-mining 'discovery' from its own dataset, such as `send-formula-coupon-to-diaper-buyer', can be a huge marketing success. Our solution will help non-IT companies replicate Google's success. Many science disciplines in NASA are typically data-intensive in nature. Many of NASA's computing environments are based on technologies 20 years ago, and thus insufficient to support growing data and computation demands. The outcome of our research will help NASA reengineering its data-intensive applications using Google's search as a blueprint, not only from user experience perspective but also from infrastructure and programming perspectives. We are aware that reinvention in this area is a high risk. Therefore, we choose to reuse proven technology and provide our innovative solutions as value-added services/libraries. By using our toolset powered by Google's engine (implemented by open-source software), NASA's scientists can do much more data analysis than just a search over a large dataset.

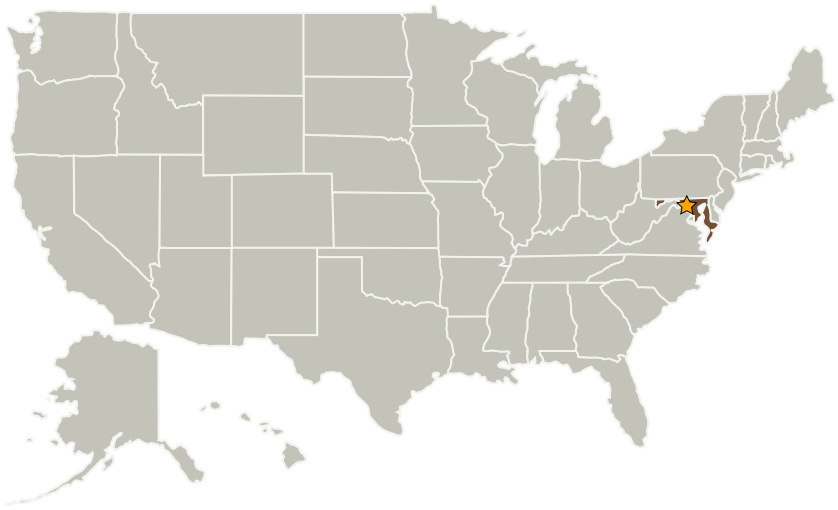
Data Intensive Scientific Computing on Petabyte Scalable Infrastructure, Phase I

Table of Contents

Data Intensive Scientific Computing on Petabyte Scalable Infrastructure, Phase I

Completed Technology Project (2009 - 2009)

Primary U.S. Work Locations and Key Partners



Organizations Performing Work	Role	Type	Location
★Goddard Space Flight Center(GSFC)	Lead Organization	NASA Center	Greenbelt, Maryland
Open Research, Inc.	Supporting Organization	Industry Minority-Owned Business, Women-Owned Small Business (WOSB)	Bethesda, Maryland

Primary U.S. Work Locations
Maryland

Organizational Responsibility

Responsible Mission Directorate:

Space Technology Mission Directorate (STMD)

Lead Center / Facility:

Goddard Space Flight Center (GSFC)

Responsible Program:

Small Business Innovation Research/Small Business Tech Transfer

Project Management

Program Director:

Jason L Kessler

Program Manager:

Carlos Torrez

Project Manager:

Ben Kobler

Principal Investigator:

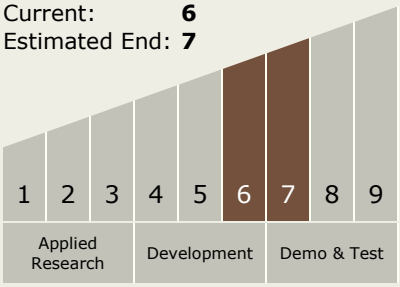
Qiming He

Data Intensive Scientific Computing on Petabyte Scalable Infrastructure, Phase I

Completed Technology Project (2009 - 2009)

Technology Maturity (TRL)

Start: **6**
Current: **6**
Estimated End: **7**



Technology Areas

Primary:

- TX11 Software, Modeling, Simulation, and Information Processing
  - TX11.4 Information Processing
    - TX11.4.2 Intelligent Data Understanding